POSTERIOR DE POLYA NO MONITORAMENTO AMOSTRAL DE PESCARIAS

Paul Gerhard KINAS [1]
Jonata Cristian WIECZYNSKI [2]

- RESUMO: Propõe-se uma abordagem bayesiana não-informativa para efetuar levantamentos amostrais de produção e de esforço em pescarias. Utiliza-se a posterior de Polya para obter inferências de parâmetros em populações finitas. A viabilidade do plano amostral aqui proposto foi analisada em um experimento piloto para coleta semanal de esforço de pesca e quantidade capturada, na pesca artesanal em Rio Grande, RS. Baseado em uma população censitária simulada contendo quatro espécies e 345 pescadores, o plano amostral foi testado utilizando uma fração amostral de $3,3\%$ de uma matriz completa com 2760 componentes. Resultados mostraram acurácias acima de 71% na estimativa da produção mensal, para todos exceto a espécie 2, a mais problemática, e aproximadamente 90% para as estimativas da produção total e do esforço para o período. O intervalo de credibilidade (ICr) teve desempenho levemente superior ao intervalo de altas densidades (HDI) em termos de cobertura; apesar de ambos terem alcançado cobertura de aproximadamente 5 pontos percentuais abaixo do valor nominal de 95%.
- PALAVRAS-CHAVE: População finita; análise bayesiana; pesca artesanal; monitoramento; amostragem

1 Introdução

A pesca profissional na modalidade artesanal é uma atividade de grande importância econômica em muitos países, pois é a fonte de renda e fator de segurança alimentar para centenas de pescadores e suas famílias. Porém, para garantir a pesca sustentável, o registro da produção e o monitoramento da atividade pesqueira,

[1]Universidade Federal do Rio Grande - FURG, Instituto de Matemática Estatística e Física, CEP: 96201-900, Rio Grande, RS, Brasil. E-mail: *paulkinas@furg.br*
[2]Universidade Federal do Rio Grande - FURG, Centro de Ciências Computacionais, PPG Computação, CEP: 96201-900, Rio Grande, RS, Brasil. E-mail: *jwieczynski@furg.br*

são fundamentais. Somente com base nas informações sobre capturas por espécie, esforço pesqueiro, preços de comercialização, custos operacionais, características dos petrechos, entre outras, que uma gestão técnica baseada em evidência pode ser implementada. Infelizmente, a disponibilidade dessas informações, ou ao menos parte delas, representa mais a exceção que a regra. Por isso é necessário viabilizar mecanismos para a coleta permanente de informações da atividade pesqueira por meio de planos amostrais de baixo custo e compatíveis com a realidade e dinâmica dessa atividade.

O desafio em produzir estatísticas pesqueiras para a pesca artesanal desenvolvida no entorno do estuário da Lagoa dos Patos (ELP) no Rio Grande do Sul, foi o elemento motivador para o presente trabalho. A grande extensão geográfica do ELP e a distribuição dispersa de comunidades de pescadores ao longo das suas margens, tornam um levantamento censitário oneroso e operacionalmente difícil. Uma alternativa aos censos é a implementação de ações pró-ativas como o automonitoramento, em que cada pescador efetua o autorregistro da sua produção; e tudo isso coordenado pelas colônias de pescadores. Embora já existam pequenas iniciativas apontando nessa direção, o automonitoramento ainda não é uma realidade no ELP. Outra possibilidade é o uso de levantamentos amostrais como forma de estimar a produção pesqueira total. Finalmente, compor automonitoramento com levantamentos amostrais é uma terceira alternativa em estudo. Nela o levantamento amostral teria a função de validar as informações censitárias obtidas por autorregistro.

Este trabalho propõe um plano amostral não-convencional, especialmente desenvolvido para atender a realidade comumente encontrada na pesca artesanal. Sua viabilidade foi testada com a implementação de um estudo piloto no município de Rio Grande. A coleta de dados ocorreu em maio de 2018 (último mês antes do período de defeso) e foi retomada de setembro até dezembro daquele ano; mostrou-se operacionalmente viável como mecanismo de aquisição de dados de produção e de esforço pesqueiros.

A inferência sobre populações finitas é normalmente feita com abordagem frequentista de amostragem (BUSSAB e BOLFARINE, 2005; COCHRAN, 1977; LUMLEY, 2010). Nesta abordagem o desenho amostral tem importância central para efetuar as inferências. No entanto, para atender as particularidades da pesca artesanal, esses planos amostrais podem tornar-se altamente complexos, dificultando a realização das inferências. Na prática, torna-se difícil (ou mesmo impossível) determinar analiticamente as margens de erro associadas aos estimadores, por exemplo. Um compromisso entre eficiência do plano amostral e viabilidade analítica se impõe nesses casos.

Uma forma de contornar tais dificuldades, é utilizar o método de Inferência Bayesiana para Populações Finitas (IBPF) (GHOSH e MEEDEN, 1997; MARTIN, 2014). Diferentemente da abordagem frequentista tradicional, e sob condições gerais de permutabilidade e ignorabilidade, definições que posteriormente tornaremos mais precisas, o plano amostral utilizado pode tornar-se irrelevante para as inferências. Isso permite empregar IBPF mesmo em situações nas quais a complexidade do

processo de coleta dos dados inviabilizaria a abordagem frequentista convencional. Além disso, ao produzir-se estimadores bayesianos, propriedades desejáveis desses estimadores (e.g. admissibilidade) são preservados.

Neste estudo apresenta-se um plano amostral viável e flexível para coleta de dados na pesca profissional artesanal, obtendo-se as estimativas populacionais por meio de IBPF.

Para atingir esse objetivo, a Metodologia foi dividida em várias subseções. Inicia-se com a notação necessária e com o desenvolvimento dos conceitos básicos que fundamentam a abordagem bayesiana de populações finitas. Em seguida, define-se a distribuição posterior de Polya como elemento central para efetuar as inferências. Na sequência, a estrutura geral do plano amostral é descrita, definindo-se também os parâmetros populacionais considerados de interesse. A construção de uma população virtual que reproduzisse aspectos típicos de uma pesca artesanal e que servisse para avaliar o desempenho da IBPF é descrita a seguir. Finalmente, na quinta e última seção da Metodologia, apresenta-se um procedimento de validação para as posteriores de Polya. As seções Resultados e Conclusões completam o texto.

2 Metodologia

2.1 Notação e Conceitos Básicos

Uma população finita com N elementos é caracterizada pelo conjunto $\boldsymbol{I} = \{I_i; i = 1, 2, \ldots, N\}$ de índices tal que $I_i = 1$ se a unidade i está incluída em uma amostra ou $I_i = 0$, caso contrário. O tamanho da população N em geral é conhecido. Os elementos i desta população são as unidades amostrais. Além do índice I_i, a cada unidade amostral estará associada uma característica quantitativa de interesse y_i (que pode ser um vetor), compondo a população objeto do estudo $\boldsymbol{Y} = \{y_i; i = 1, ..., N\}$. O interesse geralmente recai sobre parâmetros expressos na forma de funções conhecidas $T(\boldsymbol{Y})$ como o total populacional, a sua média, mediana ou desvio-padrão. No entanto este parâmetro somente poderá ser calculado se toda a população for conhecida; a isso se denomina **censo**. Se, em contraste, for extraída uma amostra $s = (I_1, \ldots, I_n)$ composta por um subconjunto de $\boldsymbol{I}$, de tamanho $n(s) \leq N$, então a população ficará subdividida no subgrupo amostrado $\boldsymbol{Y}_s = \{y_i;\ i \in s\}$ e no sub-grupo não amostrado $\boldsymbol{Y}_{s^c} = \{y_i;\ i \in s^c\}$. Ou seja, $\boldsymbol{Y} = \{\boldsymbol{Y}_s, \boldsymbol{Y}_{s^c}\}$. Já os índices $\boldsymbol{I}$ serão conhecidos para toda a população uma vez que dispomos de s. Como $T(\boldsymbol{Y}) = T(\boldsymbol{Y}_s, \boldsymbol{Y}_{s^c})$, fica claro que este parâmetro agora terá que ser estimado e que a sua estimativa passa pela inferência de $\boldsymbol{Y}_{s^c}$. Portanto, a partir apenas das características conhecidas da amostra $\boldsymbol{D} = (s, \boldsymbol{Y}_s)$, será necessário aplicar métodos de inferência para obter informações acerca das características de Y_{s^c}.

Em levantamentos amostrais frequentistas (COCHRAN, 1977; LUMLEY, 2010) as inferências sobre $T(\boldsymbol{Y})$ estão baseadas exclusivamente na distribuição de probabilidade de $\boldsymbol{I}$ que é induzida pelo particular desenho amostral utilizado na

coleta da amostra, enquanto $\boldsymbol{Y}$ são considerados valores fixos. Esta abordagem é denotada "baseada em desenho" (*design based*). Em contraste, a abordagem bayesiana (GELMAN et al., 2015) é "baseada em modelo" (*model based*) e considera $\boldsymbol{I}$ e $\boldsymbol{Y}$ aleatórios com distribuição conjunta

$$p(\boldsymbol{Y}, \boldsymbol{I}|\theta, \phi) = p(\boldsymbol{Y}|\theta) \cdot p(\boldsymbol{I}|\boldsymbol{Y}, \phi) \tag{1}$$

sendo θ e ϕ os parâmetros associados a $\boldsymbol{Y}$ e $\boldsymbol{I}$, respectivamente. Na inferência Bayesiana em populações finitas o objeto central é a distribuição preditiva posterior $p(\boldsymbol{Y}_{s^c}|\boldsymbol{Y}_s, \boldsymbol{I})$. Uma solução Bayesiana de muita relevância prática é obtida quando o desenho amostral é irrelevante para a inferência (ignorabilidade) e há permutabilidade (*exchangeability*) das unidades de $\boldsymbol{Y}$. Detalharemos as condições para que essas propriedades existam. Em seguida definimos a posterior de Polya como sendo uma distribuição preditiva não-informativa adequada para efetuar a inferência Bayesiana de $T(\boldsymbol{Y})$, quando essas propriedades forem válidas.

Usaremos $\pi(\cdot)$ para distribuições de probabilidade dos parâmetros θ e ϕ, mantendo a notação $p(\cdot)$ quando se trata de $\boldsymbol{Y}$ ou $\boldsymbol{I}$. A verossimilhança completa é dada por

$$p(\boldsymbol{Y}, \boldsymbol{I}|\theta, \phi) = p(\boldsymbol{Y}|\theta) \cdot p(\boldsymbol{I}|\boldsymbol{Y}, \phi) \tag{2}$$

enquanto a verrossimilhança observada é

$$p(\boldsymbol{Y}_s, \boldsymbol{I}|\theta, \phi) = \int p(\boldsymbol{Y}, \boldsymbol{I}|\theta, \phi) d\boldsymbol{Y}_{s^c}. \tag{3}$$

Pelo teorema de Bayes, a distribuição posterior para os parâmetros será

$$\pi(\theta, \phi|\boldsymbol{Y}_s, \boldsymbol{I}) \propto \pi(\theta, \phi) \cdot p(\boldsymbol{Y}_s, \boldsymbol{I}|\theta, \phi). \tag{4}$$

Mas, como o interesse recai sobre θ que parametriza a distribuição de $\boldsymbol{Y}$, simplesmente integramos a distribuição acima em relação a ϕ.

$$\pi(\theta|\boldsymbol{Y}_s, \boldsymbol{I}) \propto \int\int \pi(\theta)\pi(\phi|\theta)p(\boldsymbol{Y}, \boldsymbol{I}|\theta, \phi) d\boldsymbol{Y}_{s^c} d\phi. \tag{5}$$

Consideramos independência a priori entre θ e ϕ; ou seja, $\pi(\theta, \phi) = \pi(\theta) \cdot \pi(\phi)$. Também faremos a suposição $p(\boldsymbol{I}|\boldsymbol{Y}, \phi) = p(\boldsymbol{I}|\boldsymbol{Y}_s, \phi)$ que significa haver independência entre as probabilidades de inclusão de unidades amostrais e as quantidades de interesse associadas aos elementos não-observados $\boldsymbol{Y}_{s^c}$. Esta suposição torna o desenho amostral irrelevante (*ignorability*) para a inferência de $\boldsymbol{Y}_{s^c}$. Para verificar que isso de fato ocorre, reescrevemos a última expressão incluindo as suposições de independência entre os θ e ϕ e a ignorabilidade do desenho amostral.

$$\pi(\theta|\boldsymbol{Y}_s) \propto \pi(\theta) \int p(\boldsymbol{Y}|\theta) d\boldsymbol{Y}_{s^c} \int \pi(\phi) p(\boldsymbol{I}|\phi) d\phi. \tag{6}$$

sendo que a última integral se reduz a uma constante de normalização $p(\boldsymbol{I})$ e por

isso $\boldsymbol{I}$ foi removido como condicionante da distribuição posterior de θ.

Finalmente, considerando um desenho amostral ignorável, a distribuição marginal de $\boldsymbol{Y}$ é

$$p(\boldsymbol{Y}) = \int p(\boldsymbol{Y}|\theta)\pi(\theta)d\theta \tag{7}$$

Se os elementos de $\boldsymbol{Y}$ forem independentes quando condicionados a θ , então eles são permutáveis. Portanto, sob permutabilidade (*exchangeability*) de $\boldsymbol{Y}$ podemos escrever

$$p(\boldsymbol{Y}|\theta) = p(\boldsymbol{Y}_{s^c}|\theta) \cdot p(\boldsymbol{Y}_s|\theta). \tag{8}$$

com substituição na expressão anterior e notando que, $p(\boldsymbol{Y}_s|\theta)\pi(\theta) = \pi(\theta|\boldsymbol{Y}_s)p(\boldsymbol{Y}_s)$, resulta

$$\frac{p(\boldsymbol{Y})}{p(\boldsymbol{Y}_s)} = \int p(\boldsymbol{Y}_{s^c}|\theta)\pi(\theta|\boldsymbol{Y}_s)d\theta. \tag{9}$$

Mas, como

$$\frac{p(\boldsymbol{Y})}{p(\boldsymbol{Y}_s)} = p(\boldsymbol{Y}_{s^c}|\boldsymbol{Y}_s) \tag{10}$$

conclui-se que esta última integral define a distribuição preditiva de $\boldsymbol{Y}_{s^c}$ condicionado aos valores observados na amostra $\boldsymbol{Y}_s$.

Para obter uma amostra simulada desta distribuição preditiva procede-se em dois passos: primeiramente simulando valores de θ da distribuição posterior $\pi(\theta|\boldsymbol{Y}_s)$ e em seguida, simulando valores para $\boldsymbol{Y}_{s^c}$ conforme o modelo $p(\boldsymbol{Y}_{s^c}|\theta)$. Os valores de $T(\boldsymbol{Y})$ associados a essas simulações determinarão a distribuição posterior deste parâmetro populacional.

É útil dispor de uma compreensão intuitiva de permutabilidade para julgar se em alguma situação prática que se apresente, esta suposição é razoável. Seja $\boldsymbol{Y} = \{y_1, y_2, \cdots, y_N\}$ com probabilidade a priori $p(\boldsymbol{Y})$. Descobre-se que houve erro na atribuição dos índices as unidades e que estes na verdade são uma permutação $\boldsymbol{I}'$ dos índices originais $\boldsymbol{I}$; ou seja, estamos de fato falando de $\boldsymbol{Y}' = \{y_3, y_8, \cdots, y_2\}$. Se $p(\boldsymbol{Y}) = p(\boldsymbol{Y}')$ para qualquer $\boldsymbol{Y}'$ com índices permutados em relação a $\boldsymbol{I}$, então a população $\boldsymbol{Y}$ é dita permutável.

Na próxima subseção apresentaremos uma forma bastante genérica e útil para obter uma distribuição preditiva, mesmo nos casos em que há informação insuficiente para formular um modelo detalhado para $\boldsymbol{Y}$. Bastam as suposições, em geral razoáveis, de que o desenho amostral é ignorável e que os elementos de $\boldsymbol{Y}$ sejam permutáveis.

2.2 A posterior de Polya

Após observarmos $\boldsymbol{Y}_s$ queremos obter a distribuição preditiva posterior $p(\boldsymbol{Y}_{s^c}|\boldsymbol{Y}_s)$. Esta probabilidade condicional implica dependência entre as unidades amostradas e não amostradas e está implícita da suposição de permutabilidade de $\boldsymbol{Y}$. Ghosh e Meeden (1997) mostram que essa distribuição pode ser obtida facilmente

utilizando um modelo de urna de Polya e, neste caso, passando a denotar $p(\boldsymbol{Y}_{s^c}|\boldsymbol{Y}_s)$ simplesmente como a posterior de Polya.

Supondo permutabilidade para $\boldsymbol{Y}$, a sua distribuição marginal pode ser escrita como um produto de distribuições condicionalmente independentes das unidades (GELMAN et al., 2015).

$$p(\boldsymbol{Y}) = p(y_1, ..., y_N) = \int \prod_{i=1}^{N} p(y_i|\theta)\pi(\theta)d\theta \tag{11}$$

Para chegarmos a versão mais básica da distribuição de Polya, iniciamos com o modelo mais simples possível em que y_i é uma variável binária resultando em $y_i = 1$ se for sucesso e $y_i = 0$ se for fracasso. O modelo é uma distribuição Binomial com probabilidade de sucesso θ.

$$p(y_i|\theta) = \theta^{y_i}(1-\theta)^{(1-y_i)}. \tag{12}$$

Se a priori $\pi(\theta) \sim Beta(a, b)$ é uma distribuição Beta, então a distribuição marginal $p(\boldsymbol{Y})$ será Beta-binomial (GELMAN et al., 2015).Segue como propriedade desta Beta-binomial que a distribuição dos elementos não-amostrados $\boldsymbol{Y}_{s^c}$ condicionados aos amostrados $\boldsymbol{Y}_s$, será aproximadamente uma distribuição de Polya a medida que $a \to 0$ e $b \to 0$ (ver o APÊNDICE para mais detalhes).

Se $n_1 = \sum_{i=1}^{n} y_i$ é o número de sucessos na amostra e $n_2 = n - n_1$ o correspondente número de fracassos, então a distribuição de Polya é definida como

$$p(\boldsymbol{Y}_{s^c}|\boldsymbol{Y}_s) = \left[\prod_{j=1}^{2} \frac{\Gamma(n_j + r_j)}{\Gamma(n_j)}\right] \cdot \left[\frac{\Gamma(N)}{\Gamma(n)}\right]^{-1} \tag{13}$$

quando $\sum \boldsymbol{Y}_{s^c} = r_1$ e $r_2 = N - n - r_1$, sendo $p(\boldsymbol{Y}_{s^c}|\boldsymbol{Y}_s) = 0$ em caso contrário.

Se as unidades y_i forem classificadas em k categorias (com $k \leq n$), então o modelo $p(y_i|\theta)$ pode ser generalizado de Binomial para Multinomial com vetor de parâmetros $\boldsymbol{\theta} = (\theta_1, ..., \theta_k)$, tal que $\theta_i > 0$ e $\sum \theta_i = 1$. Se a priori sobre o vetor $\boldsymbol{\theta}$ for uma distribuição Dirichlet com parâmetros $(a_1, ..., a_k)$, então o resultado acima pode ser generalizado para uma distribuição de Polya com k categorias (GHOSH e MEEDEN, 1997).

Se n_j representa o número de unidades na amostra $\boldsymbol{Y}_s$ que foram classificados como pertencentes a categoria j, então a distribuição de Polya com k níveis é definida por

$$p(\boldsymbol{Y}_{s^c}|\boldsymbol{Y}_s) = \left[\prod_{j=1}^{k} \frac{\Gamma(n_j + r_j)}{\Gamma(n_j)}\right] \cdot \left[\frac{\Gamma(N)}{\Gamma(n)}\right]^{-1} \tag{14}$$

quando $\boldsymbol{Y}_{s^c}$ tem r_j unidades classificadas na categoria j com $j = 1, ..., k$, e satisfazendo $\sum_{j=1}^{k} r_j = N - n$; com $p(\boldsymbol{Y}_{s^c}|\boldsymbol{Y}_s) = 0$ caso contrário.

Apesar da aparente complexidade, é muito fácil obter amostras da distribuição

posterior de Polya. Trata-se de um processo aleatório que pode ser utilizado para efetuar as simulações. Para isso consideram-se duas urnas. A urna A contém as n unidades amostradas $\boldsymbol{Y}_s$ e a urna B contém as demais $N-n$ unidades não amostradas $\boldsymbol{Y}_{s^c}$. Sorteia-se ao acaso uma unidade de A cujo valor é representado como y_{i*} e sorteia-se simultaneamente uma unidade de B. Atribui-se o valor y_{i*} à unidade retirada de B e retorna-se ambas para a urna A. A urna A agora estará com $n+1$ unidades e a urna B com $N-n-1$. Este procedimento é repetido até que a urna B esteja vazia. Neste momento produziu-se uma amostra simulada da distribuição de Polya com n níveis (cada y_i é representante de uma categoria específica). Ao replicar este procedimento um grande número de vezes, resultará uma distribuição de Polya simulada. Finalmente, ao aplicar o cálculo da função $T(\boldsymbol{Y})$ a cada uma dessas simulações de $\boldsymbol{Y}_{s^c}$, será obtida a sua distribuição posterior. A partir dessa distribuição, todo o ferramental inferencial Bayesiano pode ser utilizado (KINAS e ANDRADE, 2010).

O pacote do *polyapost* (MEEDEN e LAZAR; GEYER, 2017) do *software* R (R CORE TEAM, 2019) permite efetuar as simulações da urna de Polya.

2.3 Plano de amostragem mensal na pesca artesanal

O plano amostral descrito a seguir utiliza uma população de tamanho N composta de pescadores artesanais cadastrados como possuidores de embarcação, podendo atuar sozinhos ou em parceria com outros pescadores. O cadastro dos pescadores artesanais com embarcação utilizado no estudo piloto no município de Rio Grande, RS foi $N = 345$.

Ao final do mês, um pescador terá as seguintes informações referentes a sua atividade pesqueira semanal:

1. o vetor dos esforços $\boldsymbol{f}_j = (f_{j1}, f_{j2}, f_{j3}, f_{j4})$, sendo f_{jw} o esforço de pesca efetuado na semana $w = 1, 2, 3, 4$ pelo pescador j;

2. os vetores $\boldsymbol{C}_{jw} = (c_{jw1}, c_{jw2}, c_{jw3}, c_{jw4})$ de capturas semanais sendo c_{jwe} a captura da espécie $e = 1, 2, 3, 4$, na semana w pelo pescador j (obs.: no estudo piloto conduzido sob este plano de amostragem foram observadas as capturas de 4 tipos de pescado (camarão, bagre, tainha e outros);

3. os vetores $\boldsymbol{U}_{jw} = (u_{jw1}, u_{jw2}, u_{jw3}, u_{jw4})$ de capturas por unidade de esforço semanais, sendo $u_{jwe} = c_{jwe}/f_{jw}$ a captura por unidade de esforço e definindo-se ainda $u_{jwe} = 0$ para $e = 1, 2, 3, 4$ quando $f_{jw} = 0$.

A cada semana são amostrados $n_w = 15$ pescadores em uma colônia diferente (escolhida por sorteio), totalizando uma amostra de tamanho $n = 60$ entrevistas no mês. Desta forma evita-se a re-amostragem de um mesmo pescador dentro do mês. Por um procedimento de amostragem sistemático, apenas uma parcela (e.g, entrevistas ímpares; $n_u = 8$) desses pescadores semanais responde simultaneamente sobre o esforço f_{jw} e as capturas $\boldsymbol{C}_{jw}$ acumulados nos últimos 7 dias; os demais (pescadores) somente informam o esforço f_{jw} para esse período. Isso porque a coleta

de informações sobre esforço é mais rápida e o propósito é ocupar o menor tempo possível dos pescadores, facilitando a cooperação em eventuais abordagens futuras. Para um pescador entrevistado por exemplo na semana $w = 2$ e cujas informações tanto de esforço como de produção foram coletados, o vetor de dados amostrados será composto pelo vetor $\boldsymbol{f}_j = (*, f_{j2}, *, *)$, seguido de $\boldsymbol{U}_j$.

Afim de efetuar as inferências com base nessa amostra, constrói-se a matriz populacional $\boldsymbol{Y}$ acrescentando $N-n$ linhas à matriz $\boldsymbol{D}_{(n\times 20)}$, tendo preenchido com *NA*s todas as células faltantes da matriz ampliada. Os *NA*s serão posteriormente substituídos utilizando o procedimento da urna de Polya conforme será descrito na Primeira Etapa para validação do estimadores (seção 2.5). Com todas as imputações concluídas, efetua-se a estimação da produção mensal de toda a frota artesanal como se esta matriz fosse, de fato, a verdadeira população $\boldsymbol{Y}$.

Inicialmente calcula-se a produção mensal de cada pescador j, sendo $\boldsymbol{Y}_j = (y_{j1}, y_{j2}, y_{j3}, y_{j4})$, e y_{je} a produção total do pescador j para a espécie e.

$$\boldsymbol{Y}_j = \sum_{w=1}^{4} \boldsymbol{U}_{jw} \cdot f_{jw} \tag{15}$$

Somando os valores de $\boldsymbol{Y}_j$, para alguma espécie e fixada, obtém-se a produção total mensal desta espécie.

$$T_e = \sum_{j=1}^{N} y_{je} \tag{16}$$

A produção total de todas as espécies somadas é:

$$T_y = \sum_{e=1}^{4} T_e \tag{17}$$

Por fim, a estimativa do esforço total do mês também pode ser calculada:

$$F = \sum_{w=1}^{4} \sum_{j=1}^{N} f_{jw} \tag{18}$$

Replicando o procedimento de amostragem de Polya um grande número de vezes, obtém-se a distribuição posterior de Polya para todos esses parâmetros. Dessas distribuições posteriores facilmente se extrai a estimativa pontual do parâmetro (e.g. usando a média ou mediana da distribuição posterior) e os percentis $2,5\%$ e $97,5\%$ para demarcar os limites do intervalo de credibilidade percentil de 95%. Os detalhamentos estão na seção 2.5.

2.4 A população virtual

Para validar a inferência via posterior de Polya para o plano amostral acima descrito, estabeleceu-se uma população virtual a partir da qual pudessem ser

extraídas amostras em conformidade com este plano. A partir dessa amostra obtém-se as inferências pertinentes que podem então ser comparadas com os parâmetros populacionais conhecidos. Seguem os detalhes pelos quais esta população foi gerada.

Para cada pescador $j = 1, 2, \ldots, N$, o número de viagens semanais (esforço) f_{jw}, com $w = 1, \ldots, 4$, são variáveis aleatórias independentes com distribuição de Poisson, $f_{jw} \sim Poisson(\mu_w)$, onde μ_w é a quantidade esperada de viagens na semana. Foram utilizados os seguintes valores: $\mu_1 = 4$, $\mu_2 = 0.5$, $\mu_3 = 1,5$ e $\mu_4 = 2,5$.

As quantidades capturadas para quatro espécies foram simuladas de modo que a presença de uma espécie pode afetar a ocorrência das outras; incorporando uma característica que ocorre na prática. Para isso induziu-se uma correlação entre as 4 espécies do seguinte modo:

- simular a **ocorrência** de cada uma das espécies em cada semana w;
- simular a **quantidade** de pescado das espécies em cada uma das f_{jw} viagens.

Primeiramente, simulou-se a ocorrência das espécies por meio de distribuições de Bernoulli. Define-se x_{jw1} a variável binária que indica a ocorrência da espécie 1 para o pescador j na semana w como sendo $x_{jw1} \sim Ber(p_1)$, onde $p_1 = 0,3$. Para a ocorrência da espécie 2 estabeleceu-se a ocorrência da espécie 1 como condicionante: $[x_{jw2}|x_{jw1} = 1] \sim Ber(p_{2|1})$ (com $p_{2|1} = 0.9$); ou $[x_{jw2}|x_{jw1} = 0] \sim Ber(p_{2|[1]})$ (com $p_{2|[1]} = 0,15$). A espécie 3 ocorre sempre que as espécies 1 e 2 não são capturadas; ou seja, $x_{jw3} = (1 - x_{jw1})(1 - x_{jw2})$. A ocorrência da espécie 4 é dependente da ocorrência da espécie 3: $[x_{jw4}|x_{jw3} = 1] \sim Ber(p_{4|3})$ (com $p_{4|3} = 0,5$); ou $[x_{jw4}|x_{jw3} = 0] \sim Ber(p_{4|[3]})$ (com $p_{4|[3]} = 0.8$).

Para simular a quantidade de pescado (em quilogramas) descarregada pelo pescador j na semana w exemplifica-se com a espécie $e = 1$; para as demais espécies o procedimento foi análogo. A quantidade capturada c_{jw1} será zero quando $x_{jw1} = 0$. Sendo $x_{jw1} = 1$, a quantidade de pescado será a soma das capturas obtidas nas f_{jw} viagens do pescador j na semana w; este valor continuará sendo zero se $f_{jw} = 0$. Por sua vez, as quantidades de pescado por viagem (i.e., por unidade de esforço), são simuladas como variáveis aleatórias independentes com distribuição qui-quadrado com parâmetro $a_e/2$, sendo a_e a captura esperada por viagem para a espécie $e = 1, 2, 3, 4$. Utilizou-se $a_1 = 40$kg, $a_2 = 5$kg, $a_3 = 100$kg e $a_4 = 15$kg. Ou seja, propõe-se espécies com quantidades médias de captura muito diferentes.

Esta população apresenta muitos desafios em termos de inferência. Há muitas capturas iguais a zero combinadas com capturas elevadas de outra(s) espécie(s). Essa característica, típica em dados de descarga pesqueira, torna muito difícil, por exemplo, utilizar-se de uma distribuição Normal multivariada para tentar modelar o vetor das capturas semanais $\boldsymbol{C}_{jw}$.

2.5 Validação dos estimadores

Dado um parâmetro populacional, deseja-se avaliar a média da distribuição posterior de Polya como estimador bem como o desempenho dos intervalos de credibilidade (ICr) e de densidade máxima (HDI). Para isso utilizou-se propriedades frequentistas de viés e de cobertura, respectivamente. Todo o procedimento é desenvolvido em duas etapas: a produção da posterior de Polya e a avaliação do seu desempenho para efetuar as inferências.

Primeira Etapa:
Extrai-se da população virtual uma amostra em conformidade com o plano amostral descrito na seção 2.3. A partir dessa amostra, as estimativas dos parâmetros populacionais são obtidas conforme segue.

- Utiliza-se várias vezes o método de Polya para completar a matriz de dados com os valores faltantes:

  1. em cada semana w preencher os $N - n_w$ valores de f_{jw};
  2. em cada semana w preencher os $N - n_u$ valores de $\boldsymbol{U}_{jw}$.

- Calculam-se os valores estimados de captura total por espécie T_e para $e = 1, 2, 3, 4$, produção total T_y e número total de viagens F para a população reconstruída a partir das amostras de Polya;

- Repete-se as etapas acima $K = 500$ vezes.

- Após as K repetições, dispõe-se das posteriores de Polya e delas se extraem os estimadores de interesse:
  (i) a média da distribuição posterior (t);
  (ii) o intervalo de credibilidade de 95% (ICr95);
  (iii) o intervalo de densidades máximas de 95% (HDI95).

Segunda Etapa:
Após repetir a primeira etapa $G = 600$ vezes, avalia-se o desempenho dos estimadores t para os seus respectivos parâmetros T por meio de:

1. Acurácia Absoluta Percentual $AAP = 100 \cdot \left(1 - \frac{|\bar{t} - T|}{T}\right)$, onde $\bar{t}$ é a média dos G valores de t;

2. As amplitudes médias dos G intervalos ICr95 e HDI95;

3. As coberturas percentuais dos intervalos de probabilidade (cobertura nominal é 95%).

3 Resultados

A matriz que compõe a população virtual (censitária) tem $N = 345$ pescadores; a cada pescador estão associados 4 componentes (escalares) contendo os esforços semanais acrescidos de outros 4 componentes (vetoriais) composto pelas descargas semanais dos 4 tipos de pescado. Do total de $345 \times 8 = 2760$ componentes da população, apenas 82 estão na amostra (15 esforços e 8 vetores de descargas semanais) o que representa uma fração amostral de apenas $3,3\%$. Todos os componentes não-amostrados são preenchidos pelo modelo da urna de Polya.

Para uma amostra extraída em conformidade com o plano amostral (seção 2.3), as estimativas dos parâmetros populacionais em forma de distribuições posteriores de Polya estão apresentadas na Figura 1. Utilizou-se 500 simulações na construção dessas posteriores.

Verifica-se (Figura 1) que todos os parâmetros populacionais são cobertos pelos intervalos de 95% exceto para a espécie 2 (T_2) em que a inferência resultou em sub-estimativa. Na população virtual esta espécie não aparece em 69.7% das capturas; quando capturadas, ela não excede 10kg em 75% dos desembarques; no entanto o valor máximo repostado é quase quatro vezes maior (36kg). Logo, em uma amostra pequena é muito comum observar zeros e valores baixos. Por outro lado, há a possibilidade de ocorrer uma sobre-estimativa se algum dos valores próximas ao extremo superior forem amostrados. Para as demais espécies, assim como para a captura total e para o esforço, o comportamento é mais estável refletindo-se em estimativas bem melhores.

Os intervalos ICr95 e HDI95 definem regiões muito similares exceto, novamente, para a espécie T_2 cuja posterior é mais fortemente assimétrica em comparação com as demais. Nos casos em que a assimetria é marcada os HDIs produzem intervalos com amplitude menores que os seus correspondentes ICrs.

Conclui-se deste resultado que, apesar da fração amostral muito reduzida, a posterior de Polya consegue obter estimativas satisfatórias para a maioria das espécies e, principalmente, para as estatísticas agrupadas (produção e esforço totais, Ty e F).

É possível, no entanto, que os resultados apresentados na Figura 1 estejam associados a uma amostra particularmente bem-sucedida. Para eliminar esta suspeita e validar o método utilizado, avaliou-se as propriedades frequentistas do procedimento inferencial por posterior de Polya conforme descrito na seção 2.5. Os resultados estão sumarizados na Tabela 1.

Apenas para o esforço F obteve-se uma acurácia superior a 90% com um valor muito próximo deste limite para a produção total T_y. Nota-se ainda que a acurácia pode variar bastante entre as espécies; característica associada a distribuição das quantidades capturadas entre pescadores e semanas. Particularmente a espécie 2, que já se mostrou problemática nos resultados apresentados na Figura 1, é confirmada aqui como a espécie cuja captura mensal estimada tem a menor acurácia; enquanto a espécie mais abundante 3 tem a melhor acurácia de todas.

Pela recomendação da *Food and Agriculture Organization of the United*

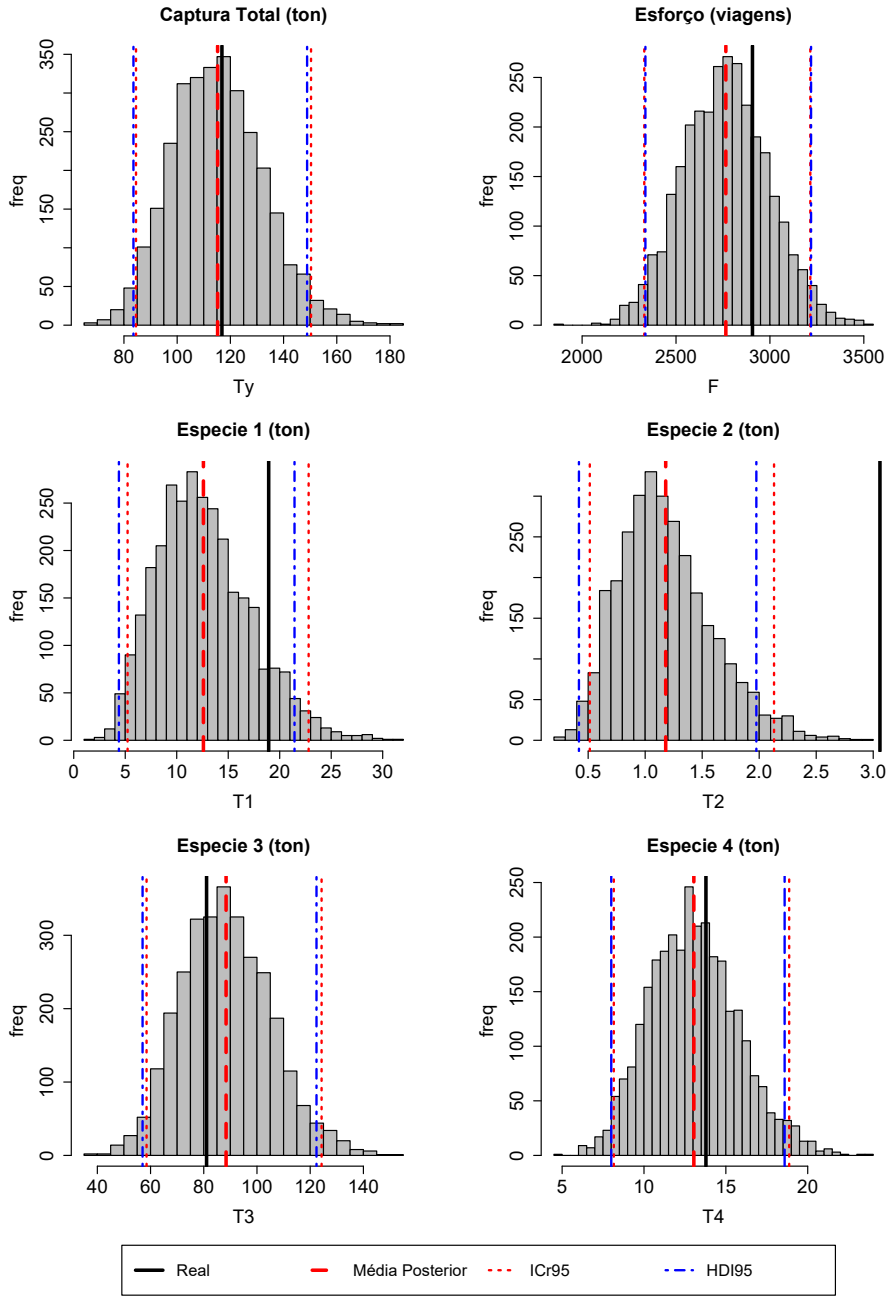

Figura 1 - Distribuições posteriores de Polya para os parâmetros populacionais.

Tabela 1 - Resultados da verificação com $G = 600$ repetições do método de Polya com $K = 500$ simulações. A média das G estimativas de Polya está na coluna $\overline{t(\boldsymbol{Y})}$ para os parâmetros viagens (F), produção total (T_y) e por espécie (T_j) para $j = 1, 2, 3, 4$

$T(\boldsymbol{Y})$	**Real**	$\overline{t(\boldsymbol{Y})}$	**Acurácia (%)**	**Amplitude**		**Cobertura (%)**	
				ICr	**HDI**	**ICr**	**HDI**
F	2907	2911,2	93,3	903,9	893,6	89,4	90,0
T_y	116,8	123,4	89,0	64,7	63,7	95,7	90,5
T_1	18,9	16,5	74,7	19,6	19,0	87,2	82,8
T_2	3,1	2,7	71,4	3,1	3,0	83,0	79,8
T_3	81,1	80,1	82,8	66,9	66,0	91,5	89,0
T_4	13,8	13,0	79,3	11,0	10,8	89,4	89,5

Nations (FAO) (STAMATOPOULOS, 2002),se o objetivo for a produção de boletins estatísticos baseados no plano de amostragem, a acurácia é aceitável se for acima de 90%. Logo, será necessário dimensionar o tamanho da amostra para alcançar esse valor mínimo, se não para todas, ao menos para as espécies eleitas como as mais importantes da pescaria.

Avaliou-se também o comportamento frequentista dos intervalos ICr e HDI quanto a sua amplitude e cobertura. Utilizou-se o valor nominal de 95% para a obtenção dos intervalos. Os HDIs por definição representam os intervalos de menor amplitude para uma dada probabilidade (KINAS e ANDRADE, 2010).Isso se confirma nas simulações em que se verifica que os HDIs sempre tem amplitude média menor que os correspondentes ICrs. Essa diferença, no entanto, é pequena nos casos examinados aqui uma vez que as posteriores de Polya são aproximadamente simétricas em torno da sua média. Por sua vez, a cobertura anunciada de 95% é atingida somente para a produção total T_y, sendo ICr superior a HDI exceto para F e T_4. Ambos tiveram o pior desempenho para a espécie 2. No geral, a cobertura de ICr mostrou-se superior ao HDI mesmo com baixa acurácia (e.g. T_1 e T_2), sugerindo maior robustez.

Antes de encerrar esta seção, algumas considerações são oportunas.

- **Unidades amostrais.** No plano amostral aqui proposto as unidades amostrais são os pescadores mestres de embarcação. Mas, dependendo da pescaria, outras unidades amostrais podem ser mais convenientes. Como, por exemplo, os locais de desembarque se forem em grande número e geograficamente dispersos; ou, os petrechos de pesca, no caso de redes de espera armados a partir da praia.

- **Esforço de pesca.** Utilizou-se, simplesmente, o número de viagens como medida de esforço. No entanto, em havendo grande variação na duração das viagens ou na dimensão dos petrechos entre pescadores, estas características

devem ser incorporadas. Sempre na intenção de aperfeiçoar ao máximo a medição atribuída a intensidade e ao poder de pesca.

- **Covariáveis.** No presente estudo ignoramos a possível presença de covariáveis. Se, por exemplo, x_i denota a potência do motor da unidade amostral i e esta informação está disponível para todo $i = 1, 2, \ldots, N$, esta informação pode ser utilizada para melhorar a predição dos esforços das unidades não amostradas, dividindo eventualmente a população em dois ou mais estratos menores. A subdivisão em estratos pode ser necessária para garantir as condições de ignorabilidade e de permutabilidade, necessárias ao método.

Conclusões

A inferência Bayesiana em populações finitas fornece uma alternativa viável para efetuar monitoramento amostral de pescarias. Para o caso da pesca artesanal, o plano amostral descrito neste trabalho é factível, conforme demostrado em um estudo piloto de campo.

Além disso, sob permutabilidade entre unidades amostradas e não-amostradas, o desenho amostral é irrelevante na construção das estimativas e de seus erros padrão. O método de simulação via urna de Polya produziu distribuições posteriores com acurácia entre 71% e 94% a partir de uma fração amostral de apenas 3.3%. Esses resultados são encorajadores uma vez que os dados de captura são desafiadores por incluírem muitos zeros associados com capturas eventualmente muito elevadas.

Os intervalos posteriores HDI, embora apresentem amplitude média inferior aos ICr correspondentes, tem cobertura inferior em 4 dos 6 casos analisados. Até que esse aspecto possa ser melhor compreendido, recomenda-se utilizar os ICr pois sugerem mais robustez. No entanto, exceto no que se refere cobertura do ICr para a captura total T_y, nos demais casos os intervalos posteriores apresentaram cobertura em torno de 5 pontos percentuais abaixo do seu valor nominal que havia sido fixado em 95%.

Agradecimentos

O primeiro autor contou com um bolsa de pesquisa no Projeto de Estatística Pesqueira financiado pelo Convênio MPA - FURG Nro. 00350.001799/2010-61. A gestora do referido projeto Liana F. Sclowitz, o técnico de campo Nilson R. Silva e o estudante de graduação Eduardo Carvalho foram essenciais para viabilizar o estudo piloto. O segundo autor utilizou-se deste estudo em seu trabalho de conclusão de curso em Matemática Aplicada. Os autores agradecem a Raquel F. Nicolette e Laura V. Miranda pelas sugestões nas versões preliminares deste manuscrito, bem como aos comentários dos revisores e editores.

KINAS, P. G.; WIECZYNSKI, J. C., Polya posterior for sample-based monitoring of fisheries. *Rev. Bras. Biom.,* Lavras, v.38, n.2, p.207-225, 2020.

- *ABSTRACT: A non-informative bayesian approach to sample-based fishery surveys is proposed. The Polya posterior for finite population parameters is used to obtain the inferences. The viability of a sampling plan was used in a pilot field experiment to collect weekly information about effort and catch from the artisanal fishery in Rio Grande, RS. Based on a simulated virtual population with four species and 345 fishermen, the sampling plan was tested using a sampling fraction of 3.3% from a complete data matrix of 2760 components. Results have shown accuracies above 71% for all but the most problematic species 2, and around 90% for estimates of total catch and cummulative effort. The percentile probability intervals (ICr) perform slightly better than the highest density interval (HDI) in terms of coverage; although both resulted about 5 percentage points bellow the nominal value of 95%.*

- *KEYWORDS: Finite population; Bayesian analysis; artisanal fishery; survey sampling.*

Referências

BUSSAB, W. O.; BOLFARINE, H. *Elementos de amostragem.* São Paulo: Blucher, 2005.

COCHRAN, W. G. *Sampling techniques.* 3.ed. New York: Wiley, 1977.

GELMAN, A. et al. *Bayesian data analysis.* 3.ed. London: Chapman & Hall/CRC, 2015.

GHOSH, M.; MEEDEN, G. *Bayesian methods for finite population sampling.* London: CRC Press, 1997.

KINAS, P. G.; ANDRADE, H. A. *Introdução à análise bayesiana (com R).* Brasil: maisQnada, 2010.

LUMLEY, T. *Complex surveys: a guide to analysis using R.* New Jersey: John Wiley & Sons, 2010.

MARTIN, R. *Bayesian analysis in finite-population models.* 2014. Disponível em: `http://homepages.math.uic.edu/~rgmartin/Teaching/Stat532/532notesnbayes.pdf`.

MEEDEN, G.; LAZAR, R.; GEYER, C. J. *polyapost: Simulating from the polya posterior.* 2017. Disponível em: hhttps://CRAN.R-project.org/package=polyaposti.

R CORE TEAM. *R: A language and environment for statistical computing.* Vienna, Austria, 2019. Disponível em: hhttps://www.R-project.orgi.

STAMATOPOULOS, C. *Sample-based fishery surveys: a technical hanbook.* Rome, Italy, Fisheries Technical Paper: FAO, 2002., 132 p.

Recebido em 26.06.2019.

Aprovado após revisão em 21.11.2019.

APÊNDICE: Derivação da distribuição de Polya

(a) Dados binários

Definimos $\boldsymbol{Y} = (y_1, ..., y_N)$ como parâmetro desconhecido cuja distribuição a priori é $p(\boldsymbol{Y})$. Após observar um subconjunto $\boldsymbol{Y}_s$ de $\boldsymbol{Y}$, queremos obter a distribuição posterior preditiva para o subconjunto não-observado $\boldsymbol{Y}_{s^c}$ denotada por $p(\boldsymbol{Y}_{s^c}|\boldsymbol{Y}_s)$. A probabilidade condicional pressupõe dependência entre valores observados e não observados e é obtida pela propriedade de permutabilidade entre os componentes de $p(\boldsymbol{Y})$. Ghosh e Meeden (1997) mostram que isso é facilmente alcançado utilizando o modelo da urna de Polya e denotam a distribuição $p(\boldsymbol{Y}_{s^c}|\boldsymbol{Y}_s)$ a distribuição posterior de Polya.

De modo genérico, uma priori para $\boldsymbol{Y}$ que seja permutável, apresenta uma simetria entre seus componentes que pode ser facilmente obtida por independência condicional entre as unidades por meio de um (super)modelo com (hiper)parâmetro θ. Isto é,

$$p(\boldsymbol{Y}) = p(y_1, ..., y_N) = \int \prod_{i=1}^{N} p(y_i|\theta)\pi(\theta)d\theta$$

Para mostrar como isso pode levar a posterior de Polya, apresentamos a situação mais simples possível em que y_i é variável binária $y = 1$ (sucesso) e $y = 0$ (fracasso), com probabilidade $p(y = 1) = \theta$ e $\theta \in [0, 1]$.

$$p(y_i|\theta) = \theta^{y_i}(1-\theta)^{1-y_i}$$

Neste caso, as incertezas sobre θ podem ser descritas convenientemente através de distribuição Beta, $\pi(\theta) \sim Beta(a, b)$.

$$\pi(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$$

Sejam $n_1 = \sum_{i\in s} y_i = \sum \boldsymbol{Y}_s$, o número de sucessos na amostra de tamanho n e $n_2 = n - n_1$ o número de fracassos. Similarmente definimos os valores r_1 e r_2 referentes ao número (desconhecido) de sucessos e de fracassos entre as $N - n$ unidades populacionais não amostradas. Então, condicionado a θ temos que

$$p(\boldsymbol{Y}|\theta) = \prod_{i=1}^{N} \theta^{y_i}(1-\theta)^{1-y_i} = \theta^{n_1+r_1}(1-\theta)^{n_2+r_2}.$$

Integrando com respeito a θ, isto é

$$p(\boldsymbol{Y}) = \int_0^1 \theta^{n_1+r_1}(1-\theta)^{n_2+r_2}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}d\theta,$$

resulta a distribuição marginal Beta-binomial para vatores $\boldsymbol{Y} = (y_1, ..., y_N)$ que

satisfazem a $T = n_1 + r_1$.

$$p(\boldsymbol{Y}) = \frac{\Gamma(a+b)\Gamma(T+a)\Gamma(N-T+b)}{\Gamma(N+a+b)\Gamma(a)\Gamma(b)}$$

Os valores de T com probabilidade positiva se restringem ao conjunto $\{0, 1, \ldots, N\}$.

Por argumento análogo, mas restrito apenas ao subconjunto amostrado $\boldsymbol{Y}_s$, resulta a distribuição Beta-binomial para $n_1 \in \{0, 1, \ldots, n\}$,

$$p(\boldsymbol{Y}_s) = \frac{\Gamma(a+b)\Gamma(n_1+a)\Gamma(n-n_1+b)}{\Gamma(n+a+b)\Gamma(a)\Gamma(b)}$$

Finalmente, pela divisão da Beta-binomial para T pela Beta-binomial para n_1 resulta a distribuição de interesse.

$$p(\boldsymbol{Y}_{s^c}|\boldsymbol{Y}_s) = \frac{p(\boldsymbol{Y})}{p(\boldsymbol{Y}_s)} = \frac{\Gamma(T+a)\Gamma(N-T+b)\Gamma(n+a+b)}{\Gamma(n_1+a)\Gamma(n-n_1+b)\Gamma(N+a+b)}$$

com probabilidades positivas para $T \in \{0, 1, \ldots, N-n\}$

A distribuição a priori para θ se manifesta nos parâmetros a e b. Quanto menores forem esses parâmetros, menor será a influência da priori sobre a distribuição preditiva posterior. Portanto, se considerarmos $a = b \approx 0$, a distribuição para o valor aleatório r_1 dependerá exclusivamente das quantidades conhecidas n_1, n e N. Desta forma chegamos a distribuição de Polya que, na sua formulação original, determina a probabilidade de obter r_1 sucessos em $R = N - n$ experimentos via urna de Polya que parte com n_1 sucessos e $n_2 = n - n_1$ fracassos (Ghosh e Meeden 1997, p.41). Ou seja,

$$p(\boldsymbol{Y}_{s^c}|\boldsymbol{Y}_s) = \left(\prod_{j=1}^{2} \frac{\Gamma(n_j + r_j)}{\Gamma(n_j)}\right) \cdot \left(\frac{\Gamma(n+R)}{\Gamma(n)}\right)^{-1}$$

É interessante registrar que com R tendendo ao infinito, a distribuição da proporção $w_R = (n_1 + r_1)/(n + R)$ de sucessos na urna de Polya converge para a distribuição $Beta(n_1, n - n_1)$.

(b) Dados em k categorias

Se, em vez de binárias, as unidades y_i forem classificadas em k categorias, sendo k algum número inteiro em $\{2, 3, \ldots, n\}$ e $\{n_1, n_2, \ldots, n_k\}$ com $n = \sum_{j=1}^{k} n_j$, as frequências de ocorrência dessas categorias na amostra de tamanho n, então o modelo Binomial se estende ao modelo Multinomial com vetor de parâmetros $\Theta = (\theta_1, ..., \theta_k)$ com as restrições $\theta_j > 0$ e $\sum \theta_j = 1$. Se a priori para o vetor Θ for a distribuição Dirichlet (uma estensão multivariada da distribuição Beta) com vetor de parâmetros $(a_1, ..., a_k)$, então o resultado da seção anterior se estende para uma

distribuição de Polya com k categorias. Isto é,

$$p(\boldsymbol{Y}_{s^c}|\boldsymbol{Y}_s) = \left(\prod_{j=1}^{k} \frac{\Gamma(n_j + r_j)}{\Gamma(n_j)}\right) \cdot \left(\frac{\Gamma(n+R)}{\Gamma(n)}\right)^{-1}$$

Neste caso, sendo $w_R^{(j)} = (n_j + r_j)/(n + R)$ definido como a proporção de unidades pertencentes a categoria j na urna de Polya, então com R tendendo ao infinito, a distribuição do vetor de proporções $(w_R^{(1)}, w_R^{(2)}, \ldots, w_R^{(k)})$ converge para a distribuição $Dirichlet(n_1, n_2, \ldots, n_k)$.